Microscaling Data Formats for Deep Learning

Year: 2023 | Citations: 107 | Authors: B. Rouhani, Ritchie Zhao, Ankit More, Mathew Hall, Alireza Khodamoradi

---

Abstract

Narrow bit-width data formats are key to reducing the computational and storage costs of modern deep learning applications. This paper evaluates Microscaling (MX) data formats that combine a per-block scaling factor with narrow floating-point and integer types for individual elements. MX formats balance the competing needs of hardware efficiency, model accuracy, and user friction. Empirical results on over two dozen benchmarks demonstrate practicality of MX data formats as a drop-in replacement for baseline FP32 for AI inference and training with low user friction. We also show the first instance of training generative language models at sub-8-bit weights, activations, and gradients with minimal accuracy loss and no modifications to the training recipe.